Regression shrinkage and grouping of highly correlated predictors with HORSES

Woncheol Jang
University of Georgia *

Johan Lim
Seoul National University †

Nicole A. Lazar
University of Georgia‡

Ji Meng Loh
AT&T Labs-Research §

Donghyeon Yu
Seoul National University ¶

February 4, 2013

Abstract

Identifying homogeneous subgroups of variables can be challenging in high dimensional data analysis with highly correlated predictors. We propose a new method called Hexagonal Operator for Regression with Shrinkage and Equality Selection, HORSES for short, that simultaneously selects positively correlated variables and identifies them as predictive clusters. This is achieved via a constrained least-squares problem with regularization that consists of a linear combination of an L_1 penalty for the coefficients and another L_1 penalty for pairwise differences of the coefficients. This specification of the penalty function encourages grouping of positively correlated predictors combined with a sparsity solution. We construct an efficient algorithm to implement the HORSES procedure. We show via simulation that the proposed

*Email:jang@uga.edu

†Email:johanlim@snu.ac.kr

‡Email: nlazar@stat.uga.edu

§Email:loh@research.att.com

¶Email:bunguji@snu.ac.kr

method outperforms other variable selection methods in terms of prediction error and parsimony. The technique is demonstrated on two data sets, a small data set from analysis of soil in Appalachia, and a high dimensional data set from a near infrared (NIR) spectroscopy study, showing the flexibility of the methodology.

Keywords and Phrases: Prediction; Regularization; Spatial correlation; Supervised clustering; Variable selection

1 Introduction

Suppose that we observe $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i = (x_{i1}, \ldots, x_{ip})^t$ is a p-dimensional predictor and y_i is the response variable. We consider a standard linear model for each of n observations

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \text{ for } i = 1, \ldots, n,$$

with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. We also assume the predictors are standardized and the response variable is centered,

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0 \text{ and } \sum_{i=1}^{n} x_{ij}^2 = 1 \quad \text{for } j = 1, \ldots, p.$$

With the dramatic increase in the amount of data collected in many fields comes a corresponding increase in the number of predictors p available in data analyses. For simpler interpretation of the underlying processes generating the data, it is often desired to have a relatively parsimonious model. It is often a challenge to identify important predictors out of the many that are available. This becomes more so when the predictors are correlated.

As a motivating example, consider a study involving near infrared (NIR) spectroscopy data measurements of cookie dough (Osborne et al., 1984). Near infrared reflectance spectral measurements were made at 700 wavelengths from 1100 to 2498 nanometers (nm) in steps of 2nm for each of 72 cookie doughs made with a standard recipe. The study aims to predict dough chemical composition using the spectral characteristics of NIR reflectance wavelength measurements. Here, the number of wavelengths p is much bigger than the sample size n.

Many methods have been developed to address this issue of high dimensionality. Section 3 contains a brief review. Most of these methods involve minimizing an objective function,

like the negative log-likelihood, subject to certain constraints, and the methods in Section 3 mainly differ in the constraints used.

In this paper, we propose a variable selection procedure that can cluster predictors using the positive correlation structure and is also applicable to data with $p > n$. The constraints we use balance between an L_1 norm of the coefficients and an L_1 norm for pairwise differences of the coefficients. We call this procedure a *Hexagonal Operator for Regression with Shrinkage and Equality Selection*, HORSES for short, because the constraint region can be represented by a hexagon. The hexagonal shape of the constraint region focuses selection of groups of predictors that are positively correlated.

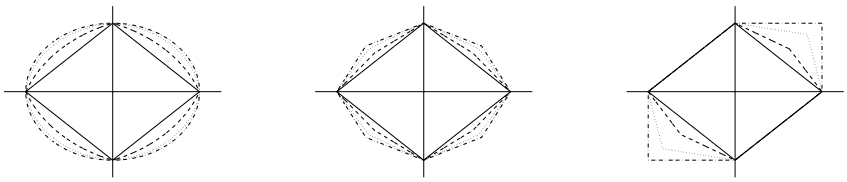
The goal is to obtain a homogeneous subgroup structure within the high dimensional predictor space. This grouping is done by focusing on spatial and/or positive correlation in the predictors, similar to supervised clustering. The benefits of our procedure are a combination of variance reduction and higher predictive power.

The remainder of the paper is organized as follows. We introduce the HORSES procedure and its geometric interpretation in Section 2. We provide an overview of some other methods in Section 3, relating our procedure with some of these methods. In Section 4 we describe the computational algorithm that we constructed to apply HORSES to data and address the issue of selection of the tuning parameters. A simulation study is presented in Section 5. Two data analyses using HORSES are presented in Section 6. We conclude the paper with discussion in Section 7.

2 Model

In this section we describe our method for variable selection for regression with *positively* correlated predictors. Our penalty terms involve a linear combination of an L_1 penalty for the coefficients and another L_1 penalty for pairwise differences of coefficients. Computation is done by solving a constrained least-squares problem. Specifically, estimates for the HORSES procedure are obtained by solving

$$\begin{aligned} \hat{\beta} &= \underset{\beta}{\operatorname{argmin}} \|y - \sum_{j=1}^{p} \beta_j x_j\|^2 \text{ subject to} \\ & \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j<k} |\beta_j - \beta_k| \leq t, \end{aligned} \tag{1}$$



(a) Elastic Net (b) OSCAR (c) HORSES

Figure 1: Graphical representation of the constraint region in the (β_1, β_2) plane for (a) Elastic Net, (b) OSCAR, and (c) HORSES

.

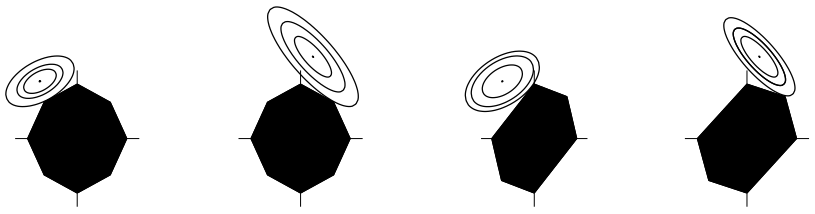
with $d^{-1} \leq \alpha \leq 1$ and d is a thresholding parameter.

As we describe in Section 3, some methods like Elastic Net and OSCAR can group correlated predictors, but they can also put negatively correlated predictors into the same group. Our method's novelty is its grouping of *positively* correlated predictors in addition to achieving a sparsity solution. Figure 1(c) shows the hexagonal shape of the constraint region induced by (1), showing schematically the tendency of the procedure to equalize coefficients only in the direction of $y = x$.

The lower bound d^{-1} of α prevents the estimates from being a solution only via the second penalty function, so the HORSES method always achieves sparsity. We recommend $d = \sqrt{p}$, where p is the number of predictors. This ensures that the constraint parameter region lies between that of the L_1 norm and of the Elastic Net method, i.e. the set of possible estimates for the HORSES procedure is a subset of that of Elastic Net. In other words, HORSES accounts for positive correlations up to the level of Elastic Net. With $d = p$, the HORSES parameter region lies within that of the OSCAR method.

In a graphical representation in the (β_1, β_2) plane, the solution is the first time the contours of the sum of squares loss function hit the constraint regions. Figure 2 gives a schematic view. Figure 2(c) shows the solution for HORSES when there is negative correlation between predictors. HORSES treats them separately by making $\hat{\beta}_1 = 0$. On the other hand, HORSES yields $\hat{\beta}_1 = \hat{\beta}_2$ when predictors are positively correlated, as in Figure 2(d).

The following theorem shows that HORSES has the exact grouping property. As the correlation between two predictors increases, the predictors are more likely to be grouped together. Our proof follows closely the proof of Theorem 1 in Bondell and Reich (2008)



(a) (b) (c) (d)

Figure 2: Graphical representation in the (β_1, β_2) plane. HORSES solutions are the first time the contours of the sum of squares function hit the hexagonal constraint region. (c) Contours centered at OLS estimate with a negative correlation. Solution occurs at $\hat{\beta}_1 = 0$; (d) Contours centered at OLS estimate with a positive correlation. Solution occurs at $\hat{\beta}_1 = \hat{\beta}_2$.

and is hence relegated to an Appendix.

Theorem 1. *Let $\lambda_1 = \lambda\alpha$ and $\lambda_2 = \lambda(1-\alpha)$ be the two tuning parameters in the HORSES criterion. Given data $\big(y, X\big)$ with centered response y and standardized predictors $X = (x_1, \ldots, x_p)^t$, let $\widehat{\beta}\big(\lambda_1, \lambda_2\big)$ be the HORSES estimate using the tuning parameters $\big(\lambda_1, \lambda_2\big)$. Let $\rho_{kl} = x_k^{\mathrm{T}} x_l$ be the sample correlation between covariates x_k and x_l.*

For a given pair of predictors x_k and x_l, suppose that both $\widehat{\beta}_k(\lambda_1, \lambda_2)$ and $\widehat{\beta}_l(\lambda_1, \lambda_2)$ are distinct from the other $\widehat{\beta}_m$. Then there exists $\lambda_0 \geq 0$ such that if $\lambda > \lambda_0$ then

$$\widehat{\beta}_k\big(\lambda_1, \lambda_2\big) = \widehat{\beta}_l\big(\lambda_1, \lambda_2\big), \quad \textit{for all } \alpha \in [d^{-1}, 1].$$

Furthermore, it must be that

$$\lambda_0 \leq \|y\| \sqrt{2(1-\rho_{kl})}/(1-\alpha).$$

The strength with which the predictors are grouped is controlled by λ_2. If λ_2 is increased, any two coefficients are morely likely to be equal. When x_i and x_j are positively correlated, Theorem 1 implies that predictors i and j will be grouped and their coefficient estimates almost identical.

3 Related work

This brief review cannot do justice to the many variable selection methods that have been developed. We highlight several of them, especially those that have links to our HORSES procedure.

While variable selection in regression is an increasingly important problem, it is also very challenging, particularly when there is a large number of highly correlated predictors. Since the important contribution of the least absolute shrinkage and selection operator (LASSO) method by Tibshirani (1996), many other methods based on regularized or penalized regression have been proposed for parsimonious model selection, particularly in high dimensions, e.g. Elastic Net, Fused LASSO, OSCAR and Group Pursuit methods (Zou and Hastie, 2005; Tibshirani et al., 2005; Bondell and Reich, 2008; Shen and Huang, 2010). Briefly, these methods involve penalization to fit a model to data, resulting in shrinkage of the estimators. Many methods have focused on addressing various possible shortcomings of the LASSO method, for instance when there is dependence or collinearity between predictors.

In the LASSO, a bound is imposed on the sum of the absolute values of the coefficients:

$$\hat{\beta} \;=\; \underset{\beta}{\operatorname{argmin}} \|y - \sum_{j=1}^{p} \beta_j x_j\|^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t,$$

where $y = (y_1, \ldots, y_n)$ and $x_j = (x_{1j}, \ldots, x_{nj})$.

The LASSO method is a shrinkage method like ridge regression (Hoerl and Kennard, 1970), with automatic variable selection. Due to the nature of the L_1 penalty term, LASSO shrinks each coefficient and selects variables simultaneously. However, a major drawback of LASSO is that if there exists collinearity among a subset of the predictors, it usually only selects one to represent the entire collinear group. Furthermore, LASSO cannot select more than n variables when $p > n$.

One possible approach is to cluster predictors based on the correlation structure and to use averages of the predictors in each cluster as new predictors. Park et al. (2007) used this approach for gene expression data analysis and introduce the concept of a *super gene*. However, it is sometimes desirable to keep all relevant predictors separate while achieving better predictive performance, rather than to use an average of the predictors. The hierarchical clustering used in Park et al. (2007) for grouping does not account for the correlation structure of the predictors.

Other penalized regression methods have also been proposed for grouped predictors (Bondell and Reich, 2008; Tibshirani et al., 2005; Zou and Hastie, 2005; Shen and Huang, 2010). All these methods except Group Pursuit work by introducing a new penalty term in addition to the L_1 penalty term of LASSO to account for correlation structure. For example, based on the fact that ridge regression tends to shrink the correlated predictors toward each other, Elastic Net (Zou and Hastie, 2005) uses a linear combination of ridge and LASSO penalties for group predictor selection and can be computed by solving the following constrained least squares optimization problem,

$$\hat{\beta} \;=\; \underset{\beta}{\operatorname{argmin}} \, ||y - \sum_{j=1}^{p} \beta_j x_j||^2 \text{ subject to } \quad \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \leq t.$$

The second term forces highly correlated predictors to be averaged while the first term leads to a sparse solution of these averaged predictors.

Bondell and Reich (2008) proposed OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression), which is defined by

$$\hat{\beta} \;=\; \underset{\beta}{\operatorname{argmin}} \, ||y - \sum_{j=1}^{p} \beta_j x_j||^2 \text{ subject to } \quad \sum_{j=1}^{p} |\beta_j| + c \sum_{j<k}^{p} \max\{|\beta_j|, |\beta_k|\} \leq t.$$

By using a pairwise L_∞ norm as the second penalty term, OSCAR encourages equality of coefficients. The constraint region for the OSCAR procedure is represented by an octagon (see Figure 1(b)). Unlike the hexagonal shape of the HORSES procedure, the octagonal shape of the constraint region allows for grouping of negatively as well as positively correlated predictors. While this is not necessarily an undesirable property, there may be instances when a separation of positively and negatively correlated predictors is preferred.

Unlike Elastic Net and OSCAR, Fused LASSO (Tibshirani et al., 2005) was introduced to account for *spatial* correlation of predictors. A key assumption in Fused LASSO is that the predictors have a certain type of ordering. Fused LASSO solves

$$\hat{\beta} \;=\; \underset{\beta}{\operatorname{argmin}} \, ||y - \sum_{j=1}^{p} \beta_j x_j||^2 \text{ subject to } \quad \sum_{j=1}^{p} |\beta_j| \leq t_1 \text{ and } \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leq t_2.$$

The second constraint, called a *fusion penalty*, encourages sparsity in the differences of coefficients. The method can theoretically be extended to multivariate data, although with a corresponding increase in computational requirements.

Note that the Fused LASSO signal approximator (FLSA) in Friedman et al. (2007) can be considered as a special case of HORSES with design matrix $X = I$. We also want to point out that our penalty function is a convex combination of the L_1 norm of the coefficients and the L_1 norm of the pairwise differenc es of coefficients. Therefore, it is not a straightforward extension of Fused LASSO in which each penalty function is constrained separately. She (2010) extended Fused LASSO by considering all possible pairwise differences and called it Clustered LASSO. However, the constraint region of Clustered LASSO does not have a hexagonal shape. As a result, Clustered LASSO does not have the *exact* grouping property of OSCAR. Consequently, She (2010) suggested to use a data-argumentation modification such as Elastic Net to achieve exact grouping.

Finally, the Group Pursuit method of Shen and Huang (2010) is a kind of supervised clustering. With a regularization parameter t and a threshold parameter λ_2, they define

$$G(z) = \begin{cases} \lambda_2 & \text{if } |z| > \lambda_2 \\ |z| & \text{otherwise,} \end{cases}$$

and estimate β using

$$\hat{\beta} \;=\; \underset{\beta}{\operatorname{argmin}} \|y - \sum_{j=1}^{p} \beta_j x_j\|^2 \text{ subject to } \quad \sum_{j<k}^{p} G(\beta_j - \beta_k) \le t.$$

HORSES is a hybrid of the Group Pursuit and Fused LASSO methods and addresses some limitations of the various methods described above. For example, OSCAR cannot handle the high-dimensional data while Elastic Net does not have the exact grouping property.

4 Computation and Tuning

A crucial component of any variable selection procedure is an efficient algorithm for its implementation. In this Section we describe how we developed such an algorithm for the HORSES procedure. The Matlab code for this algorithm is available upon request. We also discuss here the choice of optimal tuning parameters for the algorithm.

4.1 Computation

Solving the equations for the HORSES procedure (1) is equivalent to solving its Lagrangian counterpart

$$f(\beta) = \frac{1}{2}\|y - \sum_{j=1}^{p} \beta_j x_j\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j<k} |\beta_j - \beta_k|, \tag{2}$$

where $\lambda_1 = \lambda\alpha$ and $\lambda_2 = \lambda(1-\alpha)$ with $\lambda > 0$.

To solve (2) to obtain estimates for the HORSES procedure, we modify the pathwise coordinate descent algorithm of Friedman et al. (2007). The pathwise coordinate descent algorithm is an adaptation of the coordinate-wise descent algorithm for solving the 2-dimensional Fused LASSO problem with a non-separable penalty (objective) function. Our extension involves modifying the pathwise coordinate descent algorithm to solve the regression problem with a fusion penalty. As shown in Friedman et al. (2007), the proposed algorithm is much faster than a general quadratic program solver. Furthermore, it allows the HORSES procedure to run in situations where $p > n$.

Our modified pathwise coordinate descent algorithm has two steps, the descent and the fusion steps. In the descent step, we run an ordinary coordinate-wise descent procedure to sequentially update each parameter β_k given the others. The fusion step is considered when the descent step fails to improve the objective function. In the fusion step, we add an equality constraint on pairs of β_ks to take into account potential fusions and do the descent step along with the constraint. In other words, the fusion step moves given pairs of parameters together under equality constraints to improve the objective function. The details of the algorithm are as follows:

- Descent step:

  The derivative of (2) with respect to β_k given $\beta_j = \tilde{\beta}_j$, $j \neq k$, is

$$\begin{aligned}\frac{\partial f(\beta)}{\partial \beta_k} &= x_k^T x_k \beta_k - \big(y - \sum_{j\neq k} \tilde{\beta}_j x_j\big)^T x_k \\ &\quad + \lambda_1 sgn(\beta_k) + \lambda_2 \sum_{j=1}^{k-1} sgn(\tilde{\beta}_j - \beta_k) + \lambda_2 \sum_{j=k+1}^{p} sgn(\beta_k - \tilde{\beta}_j), \end{aligned} \tag{3}$$

  where the $\tilde{\beta}_j$'s are current estimates of the β_j's and $sgn(x)$ is a subgradient of $|x|$. The derivative (3) is piecewise linear in β_k with breaks at $\{0, \tilde{\beta}_j, j \neq k\}$ unless $\beta_k \notin \{0, \tilde{\beta}_j, j \neq k\}$.

– If there exists a solution to $\big(\partial f(\beta)/\partial\beta_k\big) = 0$, we can find an interval (c_1, c_2) which contains it, and further show that the solution is

$$\begin{aligned}\tilde{\beta}_k &= sgn\Big\{\tilde{y}^T x_k - \lambda_2(\sum_{j<k} s_{jk} + \sum_{j>k} s_{kj})\Big\} \\ &\quad \times \frac{\Big(\big|\tilde{y}^T x_k - \lambda_2(\sum_{j<k} s_{jk} + \sum_{j>k} s_{kj})\big| - \lambda_1\Big)_+}{x_k^T x_k},\end{aligned}$$

where $\tilde{y} = y - \sum_{j\neq k} \tilde{\beta}_j x_j$, and $s_{jk} = sgn(\tilde{\beta}_j - \frac{c_1+c_2}{2})$.

– If there is no solution to $\big(\partial f(\beta)/\partial\beta_k\big) = 0$, we let

$$\tilde{\beta}_k = \begin{cases} \tilde{\beta}_l & \text{if} \quad f(\tilde{\beta}_l) = \min\big\{f(0), f(\tilde{\beta}_j), \text{ for } j \neq k\big\} \\ 0 & \text{if} \quad f(0) \leq f(\tilde{\beta}_j), \text{ for every } j \neq k. \end{cases}$$

- Fusion step:

  If the descent step fails to improve the objective function $f(\beta)$, we consider the fusion of pairs of β_ks. For every single pair $(k, l), l \neq k$, we consider the equality constraint $\beta_k = \beta_l = \gamma$ and try a descent move in γ. The derivative of (2) with respect to γ becomes

$$\begin{aligned}\frac{\partial f(\beta)}{\partial\gamma} &= (x_k^T x_k + x_l^T x_l)\gamma - \tilde{y}^T(x_k + x_l) + 2\lambda_1 sgn(\gamma) \\ &\quad + 2\lambda_2 \textstyle\sum_{j<k,l} sgn(\tilde{\beta}_j - \gamma) + 2\lambda_2 \sum_{j>k,l} sgn(\gamma - \tilde{\beta}_j),\end{aligned}$$

  where $\tilde{y} = y - \sum_{j\neq k,l} \tilde{\beta}_j x_j$. If the optimal value of γ obtained from the descent step improves the objective function, we accept the move $\beta_k = \beta_l = \gamma$.

4.2 Choice of Tuning Parameters

Estimation of the tuning parameters α and t used in the algorithm above is very important for its successful implementation, as it is for the other methods of penalized regression. Several methods have been proposed in the literature, and any of these can be used to tune the parameters of the HORSES procedure. K-fold cross-validation (CV) randomly divides the data into K roughly equally sized and disjoint subsets D_k, $k = 1, \ldots, K$; $\bigcup_{k=1}^K D_k = \{1, 2, \ldots, n\}$. The CV error is defined by

$$\mathrm{CV}(\alpha, t) = \sum_{k=1}^{K} \sum_{i\in D_k} \left(y_i - \sum_{j=1}^{p} \widehat{\beta}_j^{(-k)}(\alpha, t) x_{ij} \right)^2,$$

where $\widehat{\beta}_j^{(-k)}(\alpha, t)$ is the estimate of β_j for a given α and t using the data set without D_k.

Generalized cross-validation (GCV) and Bayesian information criterion (BIC) (Tibshirani, 1996; Tibshirani et al., 2005; Zou et al., 2007) are other popular methods. These are defined by

$$\begin{aligned} \mathrm{GCV}(\alpha, t) &= \frac{\mathrm{RSS}(\alpha, t)}{n - \mathrm{df}}, \\ \mathrm{BIC}(\alpha, t) &= n \times \log\big(\mathrm{RSS}(\alpha, t)\big) + \log n \times \mathrm{df} \end{aligned}$$

where $\widehat{\beta}_j\big(\alpha, t\big)$ is the estimate of β_j for a given α and t, df is the degrees of freedom and

$$\mathrm{RSS}(\alpha, t) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \widehat{\beta}_j(\alpha, t) x_{ij} \right)^2.$$

Here, the degrees of freedom is a measure of model complexity. To apply these methods, one must estimate the degrees of freedom (Efron et al., 2004). Following Tibshirani et al. (2005) for Fused LASSO, we use the number of distinct groups of non-zero regression coefficients as an estimate of the degrees of freedom.

5 Simulations

We numerically compare the performance of HORSES and several other penalized methods: ridge regression, LASSO, Elastic Net, and OSCAR. We do this by generating data based on six models that differ on the number of data points n, number of predictors p, the correlation structure Σ and the true values of the coefficients β. The parameters for these six models are given in Table 1.

The first five models are very similar to those in Zou and Hastie (2005) and Bondell and Reich (2008). Specifically, the data are generated from the model

$$y = \mathrm{X}\beta + \epsilon,$$

where $\epsilon \sim N\big(0, \sigma^2\big)$. For models 1-4, we generate predictors $x_i = (x_{i1}, \ldots, x_{ip})^t$ from a multivariate normal distribution with mean 0 and covariance Σ where $\Sigma_{j,j} = 1$ for $j = 1, \ldots, p$.

Model	n	p	$\Sigma_{i,j}$	σ	β
1	20	8	$0.7^{\|i-j\|}$	3	$(3, 2, 1.5, 0, 0, 0, 0, 0)^T$
2	20	8	$0.7^{\|i-j\|}$	3	$(3, 0, 0, 1.5, 0, 0, 0, 0, 2)^T$
3	20	8	$0.7^{\|i-j\|}$	3	$(0.85, 0.85, 0.85, 0.85.0.85, 0.85, 0.85, 0.85)^T$
4	100	40	0.5	15	$(\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10})^T$
5	50	40	0.5	15	$(\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25})^T$
6	50	100	$0.7^{\|i-j\|}$	3	see text

Table 1: Parameters for the models used in the simulation study.

For model 5, the predictors are generated as follows:

$$
\begin{aligned}
x_i &= Z_1 + \eta_i^x, Z_1 \sim N(0,1), \quad i \in G_1 = \{1, \ldots, 5\} \\
x_i &= Z_2 + \eta_i^x, Z_2 \sim N(0,1), \quad i \in G_2 = \{6, \ldots, 10\} \\
x_i &= Z_3 + \eta_i^x, Z_3 \sim N(0,1), \quad i \in G_3 = \{11, \ldots, 15\} \\
x_i &\sim N(0,1), i = 16, \ldots, 40.
\end{aligned}
$$

where $\eta_i^x \sim N(0, 0.16), i = 1, \ldots, 15$. Then $\mathrm{Corr}(x_i, x_j) \approx 0.85$ for $i, j \in G_k$ for $k = 1, 2, 3$.

For model 6, we consider the scenario where $p > n$. We choose $p = 100$ because this is the maximum number of predictors that can be handled by the quadratic programming used in OSCAR. The vector of coefficients β for model 6 is given by

$$\beta = (\underbrace{3, \ldots, 3}_{5}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{5}, \underbrace{0, \ldots, 0}_{10}, \underbrace{-1.5, \ldots, -1.5}_{5}, \underbrace{0, \ldots, 0}_{10}, \underbrace{1, \ldots, 1}_{5}, \underbrace{0, \ldots, 0}_{50})^T$$

We generate 100 data sets of size $2n$ for each of the 6 models. In each data set, the final model is estimated as follows: (i) For each (α, t), we use the first n observations as a training set to estimate the model and use the other n observations as a validation set to compute the prediction error $\mathrm{PE}(\alpha, t)$; (ii) We set the tuning parameters to be the values (α^*, t^*) that minimize the prediction error $\mathrm{PE}(\alpha, t)$; (iii) The final model is estimated using the training set with $(\alpha, t) = (\alpha^*, t^*)$.

We compare the mean square error (MSE) and the model complexity of the five penalized methods. The MSE is calculated as in Tibshirani (1996) via

$$\mathrm{MSE} = (\hat{\beta} - \beta)^T \mathrm{V} (\hat{\beta} - \beta),$$

where V is the population covariance matrix for X. The model complexity is measured by the number of groups. Based on the coefficient values and correlation structure, Table 1 shows the true number of groups for each of the six scenarios. Note that the true number of groups is not always the same as the degrees of freedom. For example, we note that the true number of groups in model 5 is three based on the correlation structure although all nonzero coefficients have the same value. On the other hand, model 4 assumes a compound symmetric covariance structure, therefore the number of groups only depends on the coefficient values. Hence, the order of the coefficients does not matter and we can consider model 4 as having only one group of non-zero coefficients. We take the model complexity of model 6 to be four, based on the coefficient values. However, it is possible that some of the zero coefficients might be included as signals because of strong correlations and relatively small differences in coefficient values in this case. For example, the correlation between $\beta_{50} = 1$ and $\beta_{51} = 0$ is 0.7. Therefore it is possible that the true model complexity in this case may be bigger than four.

The simulation results are summarized in Table 2. The HORSES procedure reports the smallest dfs except for models 1 and 6. In both scenarios, the differences of df between the least complex model and HORSES is marginal (4 vs 5 in model 1 and 30 vs 33.5 in model 6). The HORSES procedure is also very competitive in the MSE comparison. Its MSE is the smallest in models 2-4 and 6 and the second or third smallest in models 1 and 5.

It is interesting to observe that HORSES is the best in model 2, but third in model 1 although the differences in MSE and df of Elastic Net and HORSES in model 1 are minor. The values of the parameters are the same in both scenarios, but variables with similar coefficients are highly correlated in model 1, while these variables have little correlation with each other in model 2. Hence we can consider the grouping of predictors as mainly determined by coefficient values in model 2 while in model 1, the correlation structure may have an important role in the grouping. This can be confirmed by comparing the median MSEs of each method in the two models 1 and 2. As expected, the median MSE in model 1 is

Table 2: The number of groups in each model used in the simulation study.

Model	1	2	3	4	5	6
Number of groups	3	3	1	1	3	4

always smaller than the median MSE in model 2. The difference in the median MSEs can be interpreted as the gain achieved by using the correlation structure when grouping. Because of the explicit form of the fusion penalty in HORSES, our procedure seems to give more weight to differences among the coefficient values while still accounting for correlations. As a result, HORSES effectively groups in model 2. Not surprisingly, the HORSES procedure is much more successful than the other procedures in finding the correct model in model 3, where it may give higher weight to the fusion penalty (α close to 1). It also has the smallest MSE among the methods. In this case, the true model is not sparse and the LASSO and Elastic Net methods fail. HORSES outperforms the other methods again in model 4. Since the model assumes the compound symmetric covariance structure, the grouping is solely based on the coefficient values. Because of the fusion penalty, the HORSES procedure is very effective in grouping and produces 3.5 as the median df while the second smallest df is 15 with OSCAR. In model 5, HORSES has the second smallest median MSE (=46.1) with Elastic Net's median MSE smallest at 40.7. However, HORSES chooses the least complex model and shows better grouping compared to Elastic Net. Model 6 considers a large p and small n case. The HORSES procedure reports the smallest MSE while the Elastic Net chooses the least complex model. However we notice that all methods report at least 30 as the df. This might be due to the fact that the true model complexity in this case is not clear, as we point out above. In summary, the HORSES procedure outperforms the other methods in choosing the least complex model and attaining the best grouping, while also providing competitive results in terms of MSE.

6 Data Analysis

6.1 Cookie dough data

In this case study, we consider the cookie dough dataset from Osborne et al. (1984), which was also analyzed by Brown et al. (2001), Griffin and Brown (2012), Caron and Doucet (2008), and Hans (2011). Brown et al. (2001) consider four components as response variables: percentage of fat, sucrose, flour and water associated with each dough piece. Following Hans (2011), we attempt to predict only the flour content of cookies with the 300 NIR reflectance measurements at equally spaced wavelengths between 1200 and 2400 nm as predictors (out of the 700 in the full data set). Also following Hans (2011) we remove the

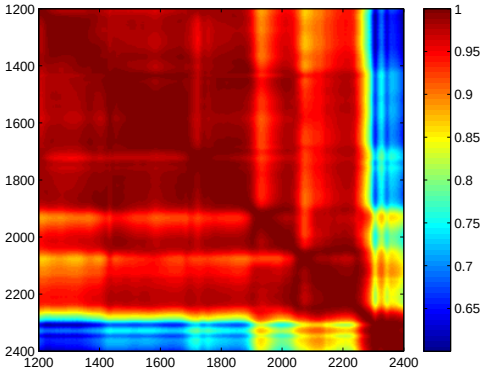
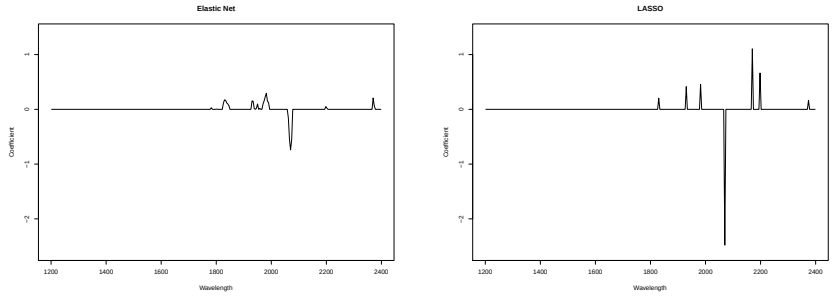


Figure 3: Graphical representation of the correlation matrix of the 300 wavelengths of the cookie dough data

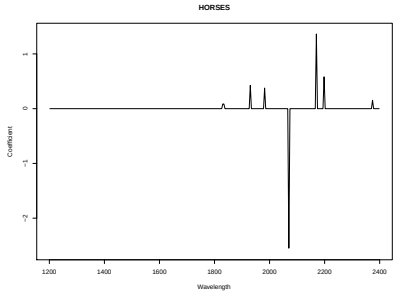
23rd and 61st observations as outliers. Then we split the dataset randomly into a training set with 39 observations and a test set with 31 observations. Figure 3 shows the correlations between NIR reflectance measurements based on all observations. There are very strong correlations between any pair of predictors in the range of 1200-2200 and 2200-2400. Note however that strong correlations do not necessarily imply strong signals in this case since the correlations can be due to the measurement errors.

With the training data set, tuning parameters of HORSES are computed to be $\alpha = 0.999$ and $\lambda = 0.1622$ (equivalently, $\lambda_1 = 0.1620$ and $\lambda_2 = 0.00016$). Since the L_1 penalty dominates the penalty function, we expect that both HORSES and LASSO will yield very similar results. We compare HORSES, LASSO and Elastic Net via the prediction mean squared error and degrees of freedoms on the test data. The OSCAR method is not included in the comparison because we are not able to apply it due to the high dimension of the data. Table 3 presents the prediction mean squared error and degrees of freedom of each method. The Elastic Net has the smallest MSE, but the differences in MSE across the three methods are small. On the other hand, the LASSO and HORSES methods provide parsimonious models with small degrees of freedom. The estimated coefficients for the LASSO, Elastic Net and HORSES methods are presented in Figure 4. Elastic Net produces 11 peaks while both LASSO and HORSES have 7 peaks. The estimated spikes from LASSO and HORSES are consistent with the results obtained in Caron and Doucet (2008). The main difference between the two methods is at wavelengths 1832 and 1836, where the LASSO estimates



(a) Elastic Net

(b) LASSO



(c) HORSES

Figure 4: Coefficient estimates for the 300 predictors of the cookie dough data

.

are 0.204 and 0 while the HORSES estimates are 0.0853 at both wavelengths. The Elastic Net has peaks at wavelength 1784 and 1804 but the other two methods do not provide a peak at those wavelengths. We observe a reverse pattern at wavelength 2176.

6.2 Appalachian Mountains Soil Data

Our next example is the Appalachian Mountains Soil Data from Bondell and Reich (2008). Figure 5 shows a graphical representation of the correlation matrix of 15 soil characteristics computed from measurements made at twenty 500-m^2 plots located in the Appalachian Mountains of North Carolina. The data were collected as part of a study on the relationship between rich-cove forest diversity and soil characteristics. Forest diversity is measured as the number of different plant species found within each plot. The values in the soil data set are averages of five equally spaced measurements taken within each plot and are standardized before the data analysis. These soil characteristics serve as predictors with forest diversity as the response.

As can be seen from Figure 5, there are several highly correlated predictors. Note that our correlation graphic shows the signed correlation values and is thus different from the

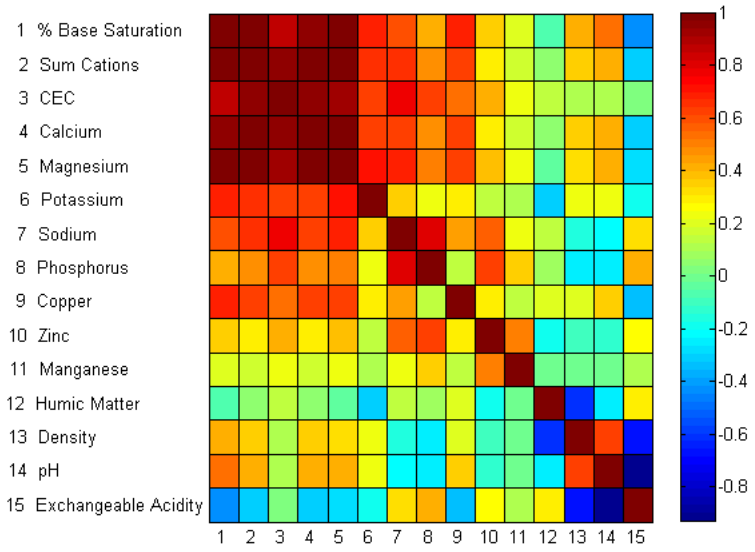


Figure 5: Graphical representation of the correlation matrix of the 15 predictors of the Appalachian soil data

one in Bondell and Reich (2008) showing the *absolute* value of correlation. The first seven covariates are closely related. Specifically they concern positively charged ions (cations). The predictors named "calcium", "magnesium", "potassium", and "sodium" are all measurements of cations of the corresponding chemical elements, while "% Base Saturation", "Sum Cations" and "CEC" (cation exchange capacity) are all summaries of cation abundance. The correlations between these seven covariates fall in the range (0.360, 0.999). There is a very strong positive correlation between percent base saturation and calcium ($r = 0.98$), but the correlation between potassium and sodium ($r = 0.36$) is not quite as high as the others. Of the remaining eight variables, the strongest negative correlation is between soil pH and exchangeable acidity ($r = -0.93$). Since both of these are measures of acidity, this appears surprising. However, exchangeable acidity measures only a subset of the acidic ions measured in pH, this subset being of more significance only at low pH values.

Note that because "Sum Cations" is the sum of the other four cation measurements the design matrix for these predictors is not full rank.

We analyze the data with the HORSES and OSCAR procedures and report the results in Table 4. Although OSCAR and HORSES use the same definition of df, the OSCAR

procedure groups predictors based on the *absolute* values of the coefficients. Therefore the number of groups is not the same as the df in OSCAR. The results for LASSO using 5-fold cross-validation and GCV can be found in Bondell and Reich (2008). The 5-fold cross-validation OSCAR and HORSES solutions are similar. They select the exact same variables, but with slightly different coefficient estimates. Since the sample size is only 20 and the number of predictors is 15, the 5-fold cross-validation method may not be the best choice for selecting tuning parameters. However, using GCV, OSCAR and HORSES provide different answers. Compared to the 5-fold cross-validation solutions, the OSCAR solution has one more predictor (% Base saturation) while the HORSES solution has 3 additional predictors (% Base saturation, Zinc, Exchangeable acidity). More interestingly, in the OSCAR solution, % Base saturation is not in the group measuring *abundance of cations*, while pH is.

On the other hand, the % Base saturation variable is included in the *abundance of cations* group. The HORSES solution also produces an additional group of variables consisting of Phosophorus and pH.

7 Conclusion

We proposed a new group variable selection procedure in regression that produces a sparse solution and also groups positively correlated variables together. We developed a modified pathwise coordinate optimization for applying the procedure to data. Our algorithm is much faster than a quadratic program solver and can handle cases with $p > n$.

Such a procedure is useful relative to other available methods in a number of ways. First, it selects groups of variables, rather than randomly selecting one variable in the group as the LASSO method does. Second, it groups positively correlated rather than both positively and negatively correlated variables. This can be useful when studying the mechanisms underlying a process, since the variables within each group behave similarly, and may indicate that they measure characteristics that affect a system through the same pathways. Third, the penalty function used ensures that the positively correlated variables do not need to be spatially close. This is particularly relevant in applications where spatial contiguity is not the only indicator of functional relation, such as brain imaging or genetics.

A simulation study comparing the HORSES procedure with ridge regression, LASSO, Elastic Net and OSCAR methods over a variety of scenarios showed its superiority in terms

of sparsity, effective grouping of predictors and MSE.

It is desirable to achieve a theoretical optimality such as the oracle property of Fan and Li (2001) in high dimensional cases. One possibility is to extend the idea of the adaptive Elastic Net (Zou and Zhang, 2009) to the HORSES procedure. Then we may consider the following penalty form:

$$\begin{aligned}\hat{\beta} &= \text{argmin}_\beta \|y - \sum_{j=1}^{p} \beta_j x_j\|^2 \text{ subject to} \\ &\alpha \sum_{j=1}^{p} \hat{w}_j |\beta_j| + (1-\alpha) \sum_{j<k} |\beta_j - \beta_k| \leq t,\end{aligned}$$

where $\hat{w}_j$ are the adaptive data-driven weights.

Investigating theoretical properties of the above estimator will be a topic of future research.

8 Appendix

Proof of Theorem 1:

Note that one can write the HORSES optimization problem in the equivalent Lagrangian form

$$\underset{\beta}{\text{argmin}} \left\{ \|y - \sum_{j=1}^{p} \beta_j x_j\|^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j<k} |\beta_j - \beta_k| \right) \right\}. \tag{4}$$

Suppose the covariates $(x_1, x_2, \ldots, x_p)$ are ordered such that their corresponding coefficient estimates satisfy

$$\widehat{\beta}_1 \leq \widehat{\beta}_2 \leq \cdots \leq \widehat{\beta}_{\text{L}} < 0 < \widehat{\beta}_{\text{L}+1} \cdots \leq \widehat{\beta}_Q$$

and $\widehat{\beta}_{Q+1} = \cdots = \widehat{\beta}_p = 0$. Let $\widehat{\theta}_1, \ldots, \widehat{\theta}_G$ denote the G unique nonzero values of the set of $\widehat{\beta}_j$, so that $G \leq Q$. For each $g = 1, 2, \ldots, G$, let

$$\mathcal{G}_g = \{j : \widehat{\beta}_j = \widehat{\theta}_g\}$$

denote the set of indices of the covariates whose estimates of regression coefficients are $\widehat{\theta}_g$. Let also $w_g = |\mathcal{G}_g|$ be the number of elements in the set $\mathcal{G}_g$

Suppose that $\widehat{\beta}_k(\lambda_1, \lambda_2) \neq \widehat{\beta}_l(\lambda_1, \lambda_2)$ and both are non-zero. In addtion, let assume $k \in \mathcal{G}_g$ and $l \in \mathcal{G}_h$ for $h > g$ without loss of generality. The differentiation of the objective function (4) with respect to β_k gives

$$- 2x_k^{\mathrm{T}}\big(y - \sum_{j=1}^{p} \widehat{\beta}_j x_j\big) + \lambda\kappa_k = 0,$$

where $u_{+,g} = \sum_{g1<g} w_{g1}$ and $u_{g,+} = \sum_{g<g2} w_{g2}$, and

$$\kappa_k = \alpha \; sgn(\widehat{\beta}_k) + \big(1-\alpha\big)\big(u_{+,g} - u_{g,+}\big). \tag{5}$$

In the same way, the differentiation of (4) with respect to β_l is

$$- 2x_l^{\mathrm{T}}\big(y - \sum_{j=1}^{p} \widehat{\beta}_j x_j\big) + \lambda\Big\{\alpha \; sgn(\widehat{\beta}_l) + \big(1-\alpha\big)\big(u_{+,h} - u_{h,+}\big)\Big\} = 0,$$

and we have, by taking their differences,

$$- 2\big(x_k^{\mathrm{T}} - x_l^{\mathrm{T}}\big)\big(y - \sum_{j=1}^{p} \widehat{\beta}_j x_j\big) + \lambda\big(\kappa_k - \kappa_l\big) = 0.$$

Since X is standardized, $\|x_k^{\mathrm{T}} - x_l^{\mathrm{T}}\|^2 = 2(1-\rho_{kl})$. This together with the fact that $\|y - X\widehat{\beta}\|^2 \leq \|y\|^2$ gives

$$|\kappa_k - \kappa_l| \leq 2\lambda^{-1}\|y\|\sqrt{2(1-\rho_{ij})}.$$

However, we find that

$$\begin{aligned} \kappa_l - \kappa_k \;\; = \;\; & \alpha\big\{sgn(\widehat{\beta}_l) - sgn(\widehat{\beta}_k)\big\} \\ & +\big(1-\alpha\big)\Big\{\big(u_{+,h} - u_{h,+}\big) - \big(u_{+,g} - u_{g,+}\big)\Big\}, \end{aligned} \tag{6}$$

is always larger than or equal to $2(1-\alpha)$. Thus, If $2\lambda^{-1}\|y\|\sqrt{2(1-\rho_{kl})} < 2(1-\alpha)$ - equivalently, $\|y\|\sqrt{2(1-\rho_{kl})}/(1-\alpha) < \lambda$ - then we encounter a contradiction.

References

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64** 115–123.

Brown, P.J., Fearn T., and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *J. Amer. Statist. Assoc.* **96** 398-408.

Caron, F. and Doucet, A.. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning. (ICML)* 88–95, Helsinki, Finland.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499.

Fan, J. and Li, R. (2001). Variable selection via non concave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R.(2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1** 302–332.

Griffin, J., and Brown, P. (2012). Bayesian hyper-lassos with non-convex penalization. *Aust. N. Z. J. Stat.* To appear. doi:10.1111/j.1467-842X.2011.00641.x.

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *J. Amer. Statist. Assoc.* **106** 1383–1393.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

Osborne, B.G., Fearn, T., Miller, A.R., and Douglas, S. (1984). Application of near infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit doughs. *J. Sci. Food Agr.* **35** 99-105.

Park, M. Y., Hastie, T., and Tibshirani, R. (2007) Averaged gene expressions for regression. *Biostatistics* **8** 212–227.

She, Y. (2010). Sparse regression with exact clustering. *Electron J. Statist.* **4** 1055-1096.

Shen, X. and Huang, H-C. (2010). Grouping pursuit through a regularization solution surface. *J. Amer. Statist. Assoc.* **105** 727–739.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc., Ser. B* **58** 267–288.

Tibshirani, R., Saunders, M., Rosset, S. Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fussed lasso. *J. Roy. Statist. Soc. Ser. B*, **67**, 91–108.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic Net. *J. Roy. Statist. Soc. Ser. B* **67** 301–320.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *Ann. Statist.* **35** 2173–2192.

Zou, H. and Zhang. H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37** 1733–1751.

Table 3: MSE and model complexity.

Case	Method	MSE Med.	MSE 10th perc.	MSE 90th perc.	DF Med.	DF 10th perc.	DF 90th perc.
C1	Ridge	2.31	0.98	4.25	8	8	8
	LASSO	1.92	0.68	4.02	5	3	8
	Elastic Net	1.64	0.49	3.26	5	3	7.5
	OSCAR	1.68	0.52	3.34	4	2	7
	HORSES	1.85	0.74	4.40	5	3	8
C2	Ridge	2.94	1.36	4.63	8	8	8
	LASSO	2.72	0.98	5.50	5	3.5	8
	Elastic Net	2.59	0.95	5.45	6	4	8
	OSCAR	2.51	0.96	5.06	5	3	8
	HORSES	2.21	1.03	4.70	5	2	8
C3	Ridge	1.48	0.56	3.39	8	8	8
	LASSO	2.94	1.39	5.34	6	4	8
	Elastic Net	2.24	1.02	4.05	7	5	8
	OSCAR	1.44	0.51	3.61	5	2	7
	HORSES	0.50	0.02	2.32	2	1	5.5
C4	Ridge	27.4	21.2	36.3	40	40	40
	LASSO	45.4	32	56.4	21	16	25
	Elastic Net	34.4	24	45.3	25	21	28
	OSCAR	25.9	19.1	38.1	15	5	19
	HORSES	21.2	19.3	33.0	3.5	1	19.5
C5	Ridge	70.2	41.8	103.6	40	40	40
	LASSO	64.7	27.6	116.5	12	9	18
	Elastic Net	40.7	17.3	94.2	17	13	25
	OSCAR	51.8	14.8	96.3	12	9	18
	HORSES	46.1	18.1	92.8	11	5.5	19.5
C6	Ridge	27.71	19.53	38.53	100	100	100
	LASSO	13.36	7.89	20.18	31	24	39.1
	Elastic Net	13.57	8.49	25.33	30	23.9	37
	OSCAR	13.16	8.56	19.16	50.00	35.9	83.7
	HORSES	12.20	7.11	22.02	33.5	24	66.3

Table 4: Biscuit dough data results

	Elastic Net	HORSES	LASSO
Mean Squared Error	2.442	2.586	2.556
Degrees of Freedom	11	7	7

Table 5: Results of analyzing the Appalachian soil data using OSCAR and HORSES, and two different methods for choosing the tuning parameters.

Variable	OSCAR (5-fold CV)	OSCAR (GCV)	HORSES (5-fold CV)	HORSES (GCV)
% Base saturation	0	-0.073	0	-0.1839
Sum cations	-0.178	-0.174	-0.1795	-0.1839
CEC	-0.178	-0.174	-0.1795	-0.1839
Calcium	-0.178	-0.174	-0.1795	-0.1839
Magnesium	0	0	0	0
Potassium	-0.178	-0.174	-0.1795	-0.1839
Sodium	0	0	0	0
Phosphorus	0.091	0.119	0.0803	0.2319
Copper	0.237	0.274	0.2532	0.3936
Zinc	0	0	0	-0.0943
Manganese	0.267	0.274	0.2709	0.3189
Humic matter	-0.541	-0.558	-0.5539	-0.6334
Density	0	0	0	0
pH	0.145	0.174	0.1276	0.2319
Exchangeable acidity	0	0	0	0.0185
Degrees of Freedom	6	5	6	7